DOCUMENT RESUME

ED 187 725 TM 800 220

AUTHOR Cleland, Winston E.; Idstein, Peter M.
TITLE In-Level versus Out-of-Level Testing of Sixth Grade Special Education Students.
PUB DATE Apr 80
NOTE 15p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Tests; Grade 6; *Grade Equivalent Scores; Intermediate Grades; Norm Referenced Tests; *Special Education; *Test Interpretation
IDENTIFIERS California Achievement Tests; New Castle County School District DE; *Out of Level Testing

ABSTRACT

When 74 sixth-grade special education students were tested on the fifth-and fourth-grade levels of the 1977 California Achievement Test, their scores dropped significantly, although converted to in-level norms. Test validity, measured by the number of students scoring at floor level, on the in-level versus below-level norms, was subtest dependent. There were no differences in the mathematics comprehension subtest, but significantly more students scored above floor level on the below-level reading vocabulary and reading comprehension subtests. A locator test, designed to aid teachers in selecting the proper test level, substantially underpredicted the percentage of students scoring above chance and floor levels. Results imply that out-of-level testing be approached with extreme caution. (CP)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED187725

# In-Level versus Out-of-Level Testing of Sixth Grade Special Education Students

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Winston E. Cleland and Peter M. Idstein

New Castle County School District

Wilmington, Delaware

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Winston E. Cleland

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM800 220

Paper presented at the Annual Meeting of the National Council of Measurement in Education, Boston, April, 1980

# In-Level versus Out-of-Level Testing of Sixth Grade Special Education Students

Winston E. Cleland and Peter M. Idstein
New Castle County School District

A common complaint from teachers who administer norm-referenced standardized tests is that the tests are inherently unfair to students who are reading below grade level. Logic dictates that if these children were tested at their functional reading level their scores would improve because they have more of an opportunity to demonstrate what they know. Unfortunately, where logic leads empiricism does not always follow.

In fact, the literature to date indicates that scores tend to drop for low-achieving students who shift from grade level to out-of-level testing. Testing students at levels other than their designated grade is called out-of-level testing.

Several studies using Title I students have shown that norm-referenced scores for students tested both out-of-level and at their designated grade level differ significantly.

Long, Schraffran and Kellogg (1977), using the Gates-MacGinitie Reading Test, found that different grade equivalent scores were obtained when Title I students received both out-of-level and in-level tests. Slaughter and Gallas (1978), using the California Achievement Test Reading Comprehension Subtest, found similar results for normal curve equivalent (NCE) scores of Title I students in the Tucson Unified School District.

Special education students represent another group of low-achieving students and there has been much pressure by special education teachers to initiate out-of-level testing. However, there have been few studies comparing in-level and out-of-level results for these students.

This study was designed to extend research in out-of-level testing to include special education students.

An additional problem created by out-of-level testing is how to choose the correct instructional level of the student. Yashida (1976) reported that teacher judgment was an accurate method of selecting the test level for special education students. Further, the test publisher of the California Achievement Tests has developed a set of locator tests that are designed to aid the teacher in selecting the proper instructional level for testing (California Achievement Tests; Test Coordinators Handbook, 1977). This study also explored the accuracy of locator test placements.

Three specific questions were investigated:

1. When special education students are tested out-of-level rather than at their designated grade level, will their norm-referenced scores (converted from out-of-level to their appropriate in-level norms) be significantly affected?
2. Will the number of special education students scoring at or below chance on in-level tests significantly decrease when these same students are tested one or two levels below their assigned grade? This question provided a test of the claim that out-of-level tests were more valid than in-level tests. The logic here is that out-of-level testing should result in significantly more students scoring above chance level.

3. Does the locator test accurately predict the correct test level for special education students?

## METHOD

The 1977 edition of the California Achievement Tests (CAT), Form C, levels 14, 15 and 16, were used in this study. The specific subtests involved were Reading Vocabulary, Reading Comprehension, and Math Computation. The conversion from out-of-level raw score to in-level NCE scores followed procedures and tables provided by CTB/McGraw-Hill, the test publisher.

Seventy-four sixth grade special education students from two elementary schools in the New Castle County School District served as subjects in this study.

Students were assigned to one of six groups via a table of random numbers. The subjects in each group received the same test levels in the same order. Group I $O_1$ took Level 16, the in-level (I) test, first and Level 15, one grade out-of-level ($O_1$), second. Group $O_1$I took Level 15 ($O_1$) first and Level 16 (I) second. The six group designations with their respective test levels and sequence are presented in Table 1.

Table 1. Test Levels and Sequence of the Six Groups

| Group | First Test | Second Test |
|---|---|---|
| I $O_1$ | Level 16 (6th grade test) | Level 15 (5th grade test) |
| $O_1$I | Level 15 (5th grade test) | Level 16 (6th grade test) |
| I $O_2$ | Level 16 (6th grade test) | Level 14 (4th grade test) |
| $O_2$I | Level 14 (4th grade test) | Level 16 (6th grade test) |
| $O_1O_2$ | Level 15 (5th grade test) | Level 14 (4th grade test) |
| $O_2O_1$ | Level 14 (4th grade test) | Level 15 (5th grade test) |

If a subject did not take both tests in a sequence, he or she was not included in the analysis. Both tests were placed in an envelope for each subject. The sequence in which the tests were to be administered was written on the front of each envelope. These envelopes, along with instructions, were given to the sixth grade teachers during a meeting at which the procedures of the study were explained in detail. The tests were then administered by the respective teachers as part of the regular District testing program. All students were also given a locator test.

The primary design was a three level ($I$, $O_1$, $O_2$) one-way analysis of variance counterbalanced for sequence effects. Question 1 was addressed by performing three ANOVA's; one each for the Reading Vocabulary, Reading Comprehension, and Math Computation subtests using NCE scores.

To examine Question 2 concerning the numbers of students scoring at chance on grade level versus out-of-level tests 2 x 3 Chi Square Analyses were used. The rows were 1) the number of students scoring at or below chance, and 2) the number of students scoring above chance. Columns 1, 2 and 3, were $I$, $O_1$, and $O_2$ respectively. Chance level was determined by dividing the number of items in a given subtest by the number of alternative choices (Slaughter and Gallas, 1978). Initial results led to completing additional 2 x 3 Chi Square Analyses using student scores at a floor level rather than at pure chance. Floor level is defined as the number of items (N) divided by the number of alternatives (A) minus one (Floor = N/A-1, see Haenn and Proctor, 1978; Roberts, 1976). The difference here is that chance level is what

a student would score by pure guessing. It is possible that chance is an unrealistic criterion in that most students do not guess 100 percent of the time.

Question 3 was addressed by comparing the number of students predicted by the locator test to be in-level with the number of students actually scoring above chance or the floor level on the in-level test.

## RESULTS

### Question 1

Table 2 presents the means and standard deviations for Reading Comprehension, Reading Vocabulary and Math Computation.

Table 2. Means and Standard Deviations for Reading Comprehension, Reading Vocabulary and Math Computation Scores

| | TEST LEVEL | | | | | |
|---|---|---|---|---|---|---|
| | I | | $O_1$ | | $O_2$ | |
| TEST | MEAN | SD | MEAN | SD | MEAN | SD |
| Reading Comprehension | 23.38 | (12.95) | 23.18 | (11.30) | 18.26 | (10.84) |
| Reading Vocabulary | 25.27 | (13.52) | 19.66 | (12.40) | 21.62 | (12.21) |
| Math Computation | 27.04 | (11.19) | 21.36 | (11.25) | 17.34 | (13.38) |

The results of the analyses of variance show significant differences for reading comprehension ($F = 3.03$, df 2/143, $p \leq .05$) and math computation ($F = 4.86$, df 2/85, $p \leq .01$). Reading vocabulary showed numerical differences but these were not statistically significant ($F = 2.44$, df 2/145, $.05 < p < .10$).

Table 3 shows the results of the analyses of variance.

Table 3. Analyses of Variance for Reading Comprehension, Reading Vocabulary and Math Computation

| TEST | SOURCE | SS | dF | MS | F |
|---|---|---|---|---|---|
| Reading Comprehension | Total | 20,420 | 145 | - | - |
| | Between groups | 830 | 2 | 415 | 3.03* |
| | Within groups | 19,590 | 143 | 137 | |
| Reading Vocabulary | Total | 24,220 | 147 | - | - |
| | Between groups | 792 | 2 | 396 | 2.44 |
| | Within groups | 23,428 | 145 | 162 | |
| Math Computation | Total | 13,756 | 87 | | |
| | Between groups | 1,408 | 2 | 704 | 4.86** |
| | Within groups | 12,348 | 85 | 145 | |

*$p < .05$
**$p < .01$

Post hoc analyses using the Duncan Multiple Range Test were performed on the Reading Comprehension and Math Computation means. Table 4 shows these data.

For Reading Comprehension $\underline{I}$ was significantly greater than $\underline{O}_2$ and $\underline{O}_1$ was significantly greater than $\underline{O}_2$. For Math Computation $\underline{I}$ was significantly greater than $\underline{O}_2$. It should be noted that $\underline{I}$ was numerically greater than $\underline{O}_1$ and $\underline{O}_2$ for Reading Vocabulary.

In all three subtests, students tested at their assigned grade level scored numerically higher than those tested below grade level. The general trend was that the further one tested out-of-level from grade six, the lower the NCE scores became, a finding consistent with

existing literature for students in grades beyond three (Long, Schraffran and Kellogg, 1977; Slaughter and Gallas, 1978).

Table 4. Duncan Multiple Range Comparisons of Difference in Reading Comprehension and Math Computation Means

| | | | | |
|---|---|---|---|---|
| I | Reading Comprehension | | | |
| | | I | $O_1$ | $O_2$ |
| | Means | 23.38 | 23.18 | 18.26 |
| I | 23.38 | - | .20 | 5.12* |
| $O_1$ | 23.18 | | - | 4.92* |
| $O_2$ | 18.26 | | | - |
| II | Math Computation | | | |
| | | I | $O_1$ | $O_2$ |
| | Means | 27.04 | 21.36 | 17.34 |
| I | 27.04 | - | 5.68 | 9.70** |
| $O_1$ | 21.36 | | - | 4.02 |
| $O_2$ | 17.34 | | | - |

*$p < .05$
**$p < .01$

Question 2

Before doing Chi Square analyses comparing students at or below chance with those above chance across test levels, ANOVA's were performed to test for order effects. There was a significant order effect ($F = 10.8$, df 1/144; $p < .05$) for reading comprehension. Consequently, the data for Chi Square analyses were taken from the first test in each sequence. There were no significant order effects for Reading Vocabulary

or Math Computation. For these subtests the data for the Chi Square Analyses were summed across test order.

The Chi Square analyses were not significant for Reading Comprehension ($\chi^2$ = .360, df 2, $p > .05$), Reading Vocabulary ($\chi^2$ = 1.55, df 2, $p > .05$) or Math Computation ($\chi^2$ = .707, df 2, $p > .05$). Table 5 shows these contingency tables. Therefore, more students did <u>not</u> score above chance level as a result of taking an out-of-level test.

Table 5. Contingency Tables for Reading Comprehension, Reading Vocabulary and Math Computation at the Chance Level

| | | TEST LEVEL | | | |
|---|---|---|---|---|---|
| | | I | $O_1$ | $O_2$ | |
| I | Reading Comprehension | | | | |
| | Chance | 6 | 5 | 5 | 16 |
| | Above Chance | 17 | 19 | 21 | 57 |
| | | 23 | 24 | 26 | 73 |
| II | Reading Vocabulary | | | | |
| | Chance | 17 | 12 | 12 | 41 |
| | Above Chance | 31 | 38 | 38 | 107 |
| | | 48 | 50 | 50 | 148 |
| III | Math Computation | | | | |
| | Chance | 2 | 2 | 4 | 8 |
| | Above Chance | 26 | 26 | 28 | 80 |
| | | 28 | 28 | 32 | 88 |

This finding was a surprise. In an attempt to better understand these data, we examined what would happen to students scoring at a "floor" level rather than at a "chance" level. Table 6 shows these data. The Chi Square analyses were significant for Reading Comprehension ($\chi^2$ = 6.45, df 2, $p < .05$) and Reading Vocabulary ($\chi^2$ = 10.83, df 2, $p < .01$)

but not for Math Computation ($\chi^2$ = 1.89, df 2, $p > .05$). Significantly fewer students scored at the floor level on the out-of-level tests than on the in-level tests for Reading Comprehension and Reading Vocabulary. For Math Computation there was no difference between the number of students scoring at the floor level on the in-level test and the number scoring at the floor level on the out-of-level test.

Table 6. Contingency Tables for Reading Comprehension, Reading Vocabulary and Math Computation at the Floor Level versus above Floor Level

| | | TEST LEVEL | | | |
|---|---|---|---|---|---|
| | | I | $O_1$ | $O_2$ | |
| I | Reading Comprehension | | | | |
| | Floor | 17 | 9 | 13 | 39 |
| | Above Floor | 6 | 15 | 13 | 34 |
| | | 23 | 24 | 26 | 73 |
| II | Reading Vocabulary | | | | |
| | Floor | 28 | 34 | 18 | 80 |
| | Above Floor | 20 | 16 | 32 | 68 |
| | | 48 | 50 | 50 | 148 |
| III | Math Computation | | | | |
| | Floor | 6 | 7 | 7 | 20 |
| | Above Floor | 22 | 21 | 25 | 68 |
| | | 28 | 28 | 32 | 88 |

Thus, depending upon the criterion and subtest one chooses, these data may or may not support claims of increased validity via out-of-level testing. For the Math Computation subtest, there were no differences at either floor or chance levels; thus these data did not support the claim of increased validity. However, for the two reading subtests

significantly more students scored above the floor level in the out-of-level tests, thus lending some support to the claim of increased validity for out-of-level testing.

Question 3

Table 7 compares the percentage of students predicted by the locator test to be in-level with the percentage of students actually scoring above the chance and floor levels of each subtest. In all cases, the percentage of students scoring above the chance and floor levels was substantially greater than the percentage of students predicted by the locator test to be in-level. The locator test might be underpredicting the appropriate test level for these students.

Table 7. The Percentage of Students Predicted by the Locator Test to be In-Level versus the Actual Percentage of Students Scoring Above Chance and Floor Levels

| Test | In-Level Prediction Locator Test (%) | Actual Results | |
|---|---|---|---|
| | | Above Chance (%) | Above Floor (%) |
| Reading Comprehension | 9.1 | 77.3 | 22.7 |
| Reading Vocabulary | 8.7 | 63.0 | 39.1 |
| Math Computation | 11.5 | 88.5 | 80.8 |

## CONCLUSIONS

### Question 1

These findings demonstrate that when special education students are tested out-of-level their norm-referenced scores (converted from out-of-level to their appropriate in-level norms) significantly drop. The common opinion held by special education teachers and administrators is that if their students were tested one or two levels below their designated grade their norm-referenced scores would improve because they would have more of an opportunity to demonstrate what they know. The data presented in this study fail to support this opinion. Instead, they support the counter position that special education students' norm-referenced scores will drop if an out-of-level testing program is initiated. The possible detrimental effects of depressed scores on the self-concept of these students and the political implications of lower test scores should be seriously considered before initiating an out-of-level testing program.

### Question 2

The number of students scoring above the chance or floor levels on norm-referenced tests may or may not increase when they are given out-of-level tests. The data appear to be subtest dependent. These findings question the claim that out-of-level testing provides more valid results than in-level testing. These data further question the claim that out-of-level testing provides better diagnostic information.

### Question 3

The locator test may underestimate the appropriate test level for special education students. Additional research specifically designed

to explore the best methods of identifying the most appropriate test levels for special education students seems warranted.

One criticism of this study is that one or two levels below grade level is not sufficient for special education students. It is possible that a majority of sixth-grade special education students are reading more than two years below grade level. However true this may be, the point may be moot. Ross Green, Director of Research for CTB/McGraw-Hill, has stated that two levels down is as far as one can go with the CAT and still maintain the instructional integrity of the tests. Below two levels, the instructional bridge to the curriculum may be compromised (personal communication). However, Plake and Hoover (1979), using the Iowa Tests of Basic Skills and a contrasting item difficulty method, present data to support their argument that norm-referenced differences (G.E. scores) were not caused by differences in the applicability of items to the curriculum. Green's position is still open to question.

It is possible that the differences between norm-referenced scores on in-level and out-of-level tests are due to the statistical procedures for converting out-of-level raw scores to in-level norm-referenced scores. We echo the plea of Slaughter and Gallas (1978) that empirical norms should be developed for out-of-level testing by testing students at multiple test levels in the norming procedure. Perhaps norms should be developed for special education students.

In any event, it is clear from these results as well as others cited that any decision to test out-of-level should be approached with extreme caution.

## REFERENCES


California Achievement Tests: Test Coordinators Handbook. CTB/McGraw-Hill, Del Monte Research Park, Monterey, California, 1977.

Haenn, Joseph F. and Proctor, Diane C. A practitioner's guide to out-of-level testing. A paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March, 1978.

Long, J., Schraffran, J. and Kellogg, T. Effects of out-of-level survey testing on reading achievement scores of Title I, ESEA students. Journal of Educational Measurement, 1977, 14, pp. 203-213.

Plake, B.S. and Hoover, H.D. The comparability of equal raw scores obtained from in-level and out-of-level testing: One source of the discrepancy between in-level and out-of-level grade equivalent scores. Journal of Educational Measurement, 1979, 16, pp. 279-284.

Roberts, A.O.H. Out of level testing - ESEA Title I Evaluation and Reporting System. Technical Paper No. 6. Mountain View, California. RMC Research Corporation. October, 1976.

Slaughter, H.B. and Gallas, E.J. Will out-of-level norm-referenced testing improve the selection of program participants and the diagnosis of reading comprehension in ESEA Title I programs? Paper presented at the Annual Meeting of the American Educational Research Association, Toronto, Canada, March 27-31, 1978.

Yashida, R.K. Out-of-level testing of special education students with a standardized achievement battery. Journal of Educational Measurement, 1976, 13, pp. 215-221.